Sequential Compressed Sensing

Dmitry Malioutov

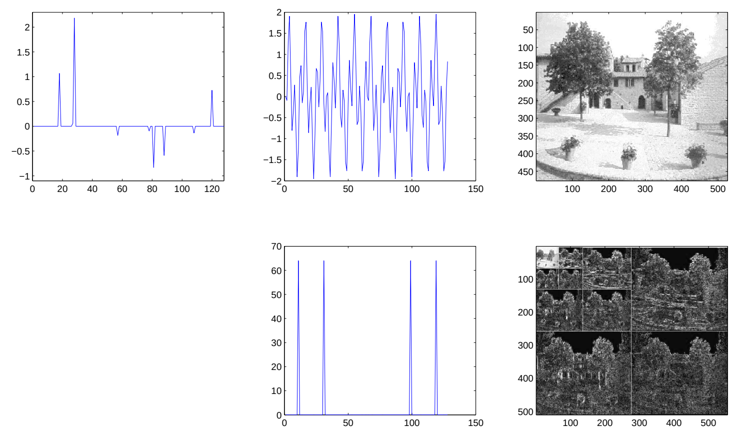
DRW, research done mostly at MIT

Joint work with Sujay Sanghavi and Alan Willsky

September 03, 2009

Motivation

Many important classes of signals are either sparse, or compressible.
Examples: sparsity in : (a) standard, (b) Fourier, (c) wavelet basis.



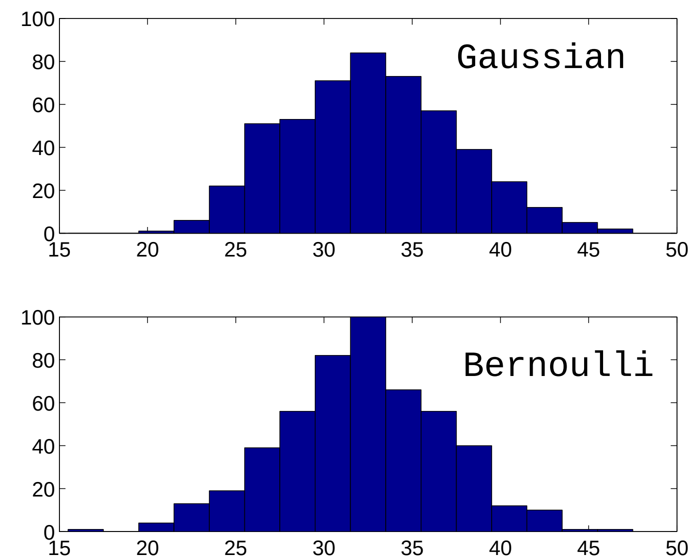
CS: K-sparse $\mathbf{x} \in \mathbb{R}^N$. We take $M << N$ measurements $y = A\mathbf{x} + \mathbf{n}$, and try to recover $\mathbf{x}$ *knowing* that it is sparse.

Related problems: recovering structure of graphical models from samples, recovering low-rank matrices from few measurements, ...

Motivation

CS: for certain random A, $\mathbf{x}$ can be efficiently recovered with high prob. after $O(K\log(N/K))$ samples, where $\mathbf{x}$ is K-sparse.

Req. M for signal with $K = 10$.



However:

- may not know K a-priori
- such bounds are not available for all decoders
- constants may not be tight.

How many samples to get?

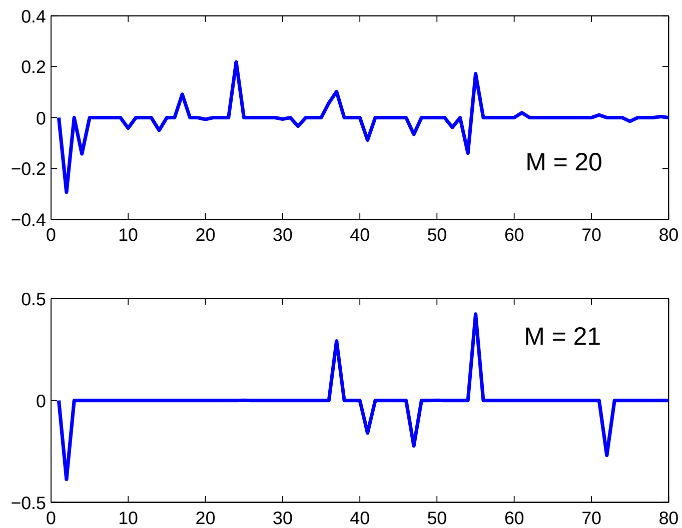
Our approach: receive samples sequentially $y_i = \mathbf{a}_i'\mathbf{x}$ and stop once we know that enough samples have been received.

Presentation Outline

1. CS formulation with sequential observations.
2. Stopping rule for the Gaussian case.
3. Stopping rule for the Bernoulli case.
4. Near-sparse and noisy signals.
5. Efficient solution of the sequential problem.

Batch CS

Batch CS: suppose $\mathbf{y} = A\mathbf{x}^*$. Find the sparsest $\mathbf{x}$ satisfying $\mathbf{y} = A\mathbf{x}$. Relaxations: greedy methods, convex ℓ_1, non-convex ℓ_p, sparse Bayesian learning, message passing, e.t.c. – these all give sparse solutions. How to verify that the solution also recovers $\mathbf{x}^*$?



Top plot: reconstruction from $M = 20$ samples, $N = 80$.

Bottom plot: using $M = 21$ samples (correct).

To guarantee correct reconstruction with high probability - we need a superfluous number of samples to be on the 'safe side'.

Sequential CS formulation

Observations are available in sequence: $y_i = \mathbf{a}_i'\mathbf{x}^*$, $i = 1, .., M$.

At step M we use any sparse decoder to get a feasible solution $\hat{\mathbf{x}}_M$, e.g. the ℓ_1 decoder:

$$\hat{\mathbf{x}}_M = \arg\min ||\mathbf{x}||_1 \quad \text{s.t.} \quad \mathbf{a}_i'\mathbf{x} = y_i, \quad i = 1, .., M$$

and either declare victory and stop, or ask for another sample.

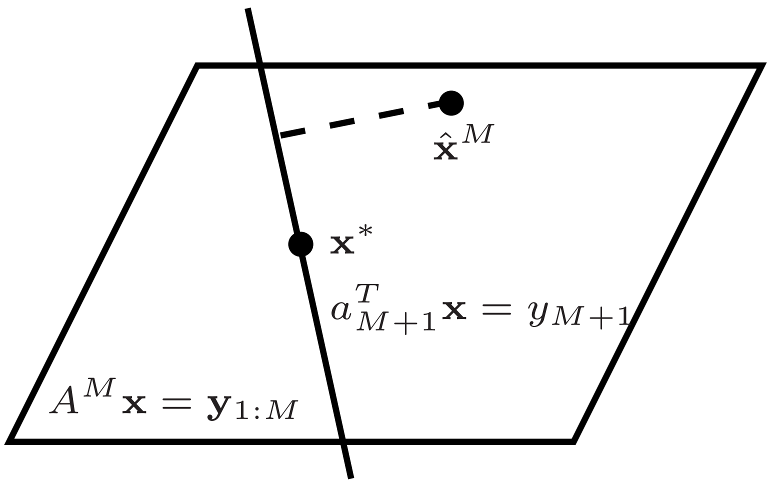
Q: How does one know when enough samples have been received?

Waiting for $M \propto CK \log(N/K)$: requires knowledge of K, $K = \|\mathbf{x}^*\|_0$. Also only rough bounds on proportionality constants may be known, and not even for all algorithms.

Gaussian measurement case

Receive $y_i = \mathbf{a}_i'\mathbf{x}^*$, where $\mathbf{a}_i \sim \mathcal{N}(0, I)$ i.i.d. Gaussian samples.

Claim: if $\hat{\mathbf{x}}^{M+1} = \hat{\mathbf{x}}^M$ then $\hat{\mathbf{x}}^M = \mathbf{x}^*$ with probability 1.
$A^M \triangleq [\mathbf{a}_1', ...\mathbf{a}_M']'$

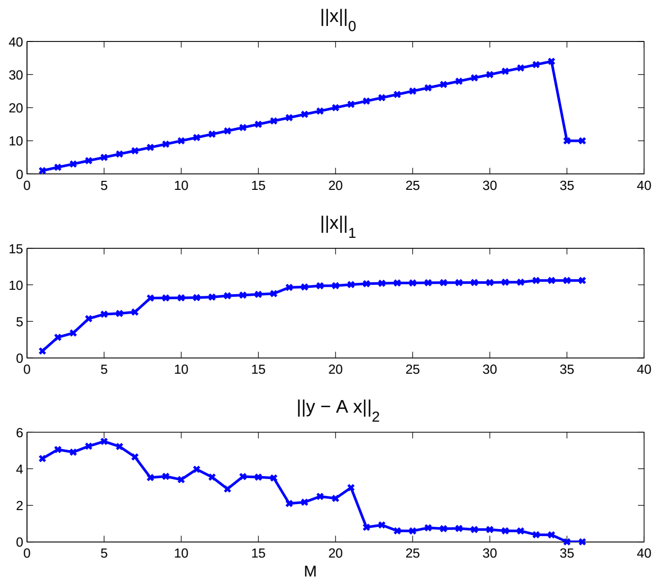


A new sample $\mathbf{a}_{M+1}'\mathbf{x} = y_{M+1}$ passes a random hyperplane through $\mathbf{x}^*$. Probability that this hyperplane also goes through $\hat{\mathbf{x}}^M$ is zero.

Gaussian case (continued)

Even simpler rules: (i) if $\|\hat{\mathbf{x}}^M\|_0 < M$ or if (ii) $\mathbf{a}'_{M+1}\hat{\mathbf{x}}^M = y_{M+1}$ then $\hat{\mathbf{x}}^M = \mathbf{x}^*$.

This works because for a random Gaussian matrix all $M \times M$ submatrices are non-singular with prob. 1.



Example: $N = 100$, and $K = 10$.

Top plot: $\|\hat{\mathbf{x}}^M\|_0$.
Midle plot: $\|\hat{\mathbf{x}}^M\|_1$.
Bottom plot: $\|\mathbf{x}^* - \hat{\mathbf{x}}^M\|_2$.

Bernoulli case

Let $\mathbf{a}_i$ have equiprobable i.i.d. Bernoulli entries ± 1. Now $M \times M$ submatrices of A^M can be singular (non-0 probability).

The stopping rule for the Gaussian case does not hold. We modify it as follows: wait until $\hat{\mathbf{x}}^M = \hat{\mathbf{x}}^{M+1} = ... = \hat{\mathbf{x}}^{M+T}$.

Claim: After T-step agreement $P(\hat{\mathbf{x}}^{M+T} \neq \mathbf{x}^*) < 2^{-T}$.

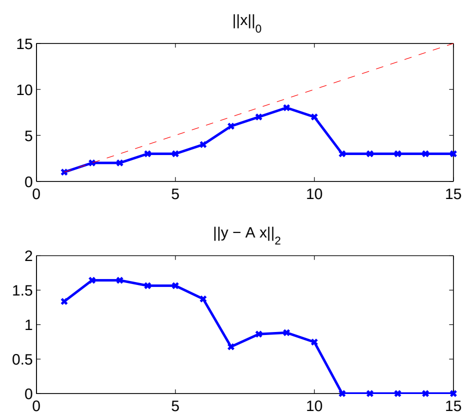
Proof depends on Lemma (Tao and Vu): Let $\mathbf{a} \in \{-1, 1\}^N$ be an i.i.d. equiprobable Bernoulli. Let W be a fixed d-dimensional subspace of $\mathbb{R}^N$, $0 \leq d < N$. Then $P(\mathbf{a} \in W) \leq 2^{d-N}$.

Suppose $\hat{\mathbf{x}}^M \neq \mathbf{x}^*$. Let $\mathcal{J}$ and $\mathcal{I}$ be their supports, $L = |\mathcal{I} \cup \mathcal{J}|$. Then $\mathcal{A} = \{\mathbf{a}_{\mathcal{I}\cup\mathcal{J}} \mid (\hat{\mathbf{x}}^M - \mathbf{x}^*)' \, \mathbf{a}_{\mathcal{I}\cup\mathcal{J}} = 0\}$ is an $(L-1)$-dim. subspace of $\mathbb{R}^L$. Prob that $\mathbf{a}^{M+1}_{\mathcal{I}\cup\mathcal{J}}$ belongs to $\mathcal{A}$ is at most $1/2$. $\diamond$

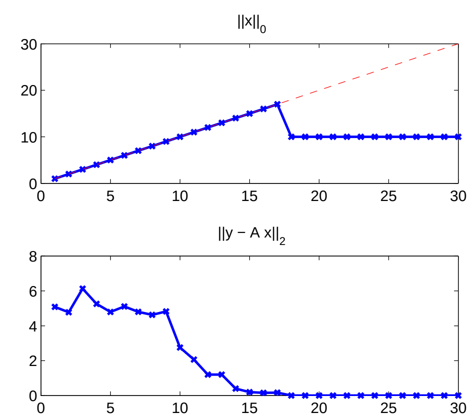
Bernoulli case (continued)

Rule only uses T. Ideally we should also use M and N: errors are more likely for smaller M and N.

Conjecture: for $M \times M$ matrix $P(\det(A) = 0) \propto M^2 2^{1-M}$. (Main failure: a pair of equal rows or columns). Best provable upper bound is still quite loose. Such analysis could allow shorter delay.



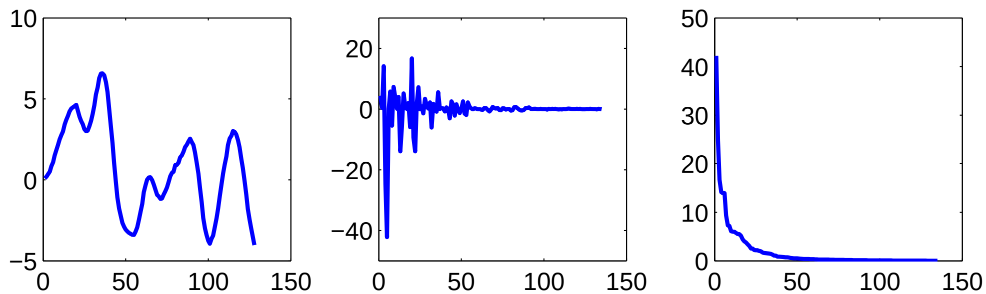
Example with $K = 3$, $N = 40$.



Example with $K = 10$, $N = 40$.

Near-sparse signals

In many practical settings signals are near-sparse: e.g. Fourier or wavelet transforms of smooth signals.



(a) signal,
(b) wav. coeffs.,
(c) coeffs. sorted.

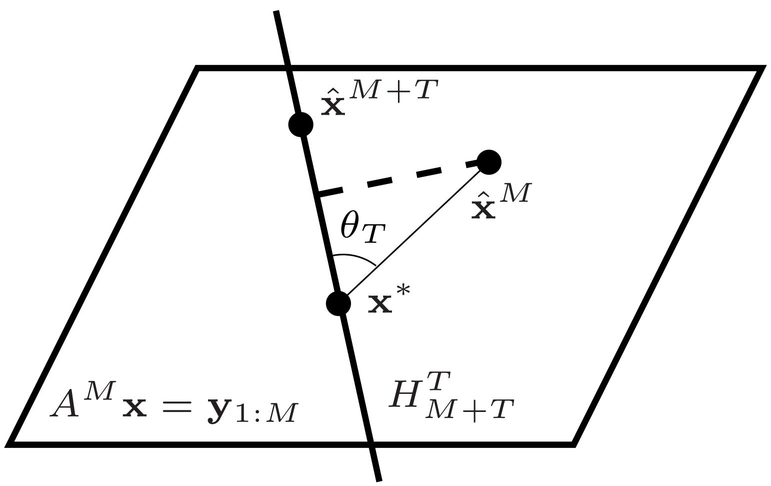
CS results: with roughly $O(K \log N)$ samples, $\hat{\mathbf{x}}^M$ has similar error to keeping K largest entries in $\mathbf{x}^*$.

Our approach:
Given $\hat{\mathbf{x}}^M$, we obtain T new samples, and find distance from $\hat{\mathbf{x}}^M$ to $H_{M+T} \triangleq \{x \mid y_i = \mathbf{a}_i' x,\ 1 \leq i \leq M+T\}$. This distance can be used to bound the reconstruction error $\|\mathbf{x}^* - \hat{\mathbf{x}}^M\|_2$.

Near-sparse signals (continued)

Let $H_{M+T} \triangleq \{\mathbf{x} \mid y_i = \mathbf{a}_i'\mathbf{x},\ i = 1, .., M+T\}$. Let θ_T be the angle between the line $(\mathbf{x}^*, \hat{\mathbf{x}}^M)$ and H_{M+T}.

$$d(\mathbf{x}^*, \hat{\mathbf{x}}^M) = \frac{d(\hat{\mathbf{x}}^M, H_{M+T})}{\sin(\theta_T)} \triangleq C_T\ d(\hat{\mathbf{x}}^M, H_{M+T})$$

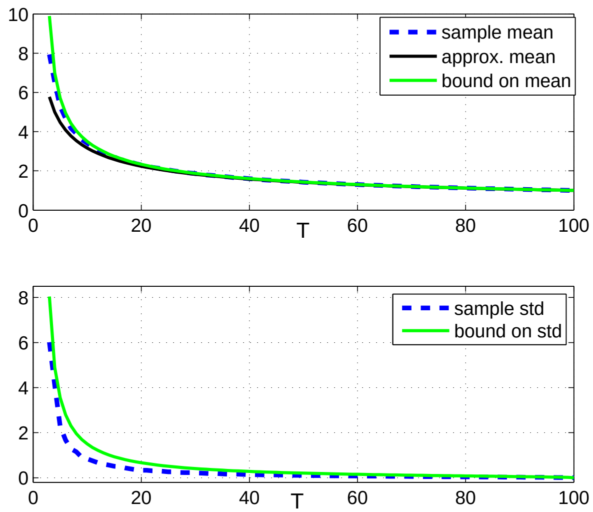


Let $L = N - M$.
Using properties of χ_L, χ^2_L and Jensen's ineq. we have:

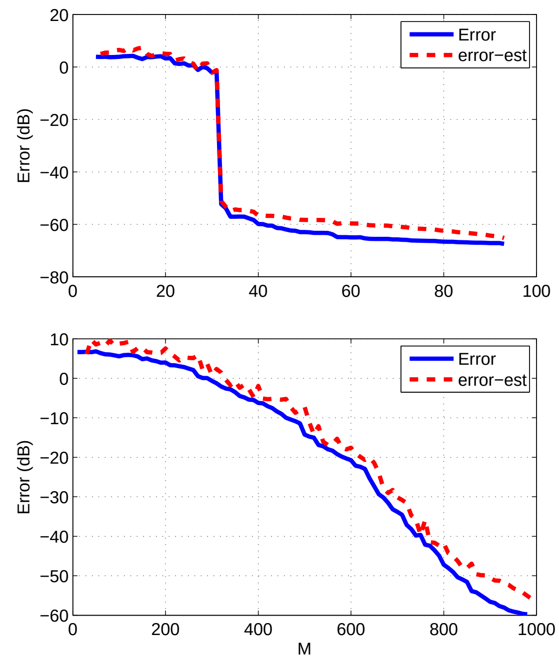
$E[\frac{1}{\sin(\theta)}] \approx \sqrt{\frac{L}{T}} \leq \sqrt{\frac{L-2}{T-2}}$

$Var\left[\frac{1}{\sin(\theta)}\right] \leq \frac{L-2}{T-2} - \frac{L}{T}$

Examples: near-sparse signals



(Top) sample C_T, approx and bound.
(Bottom) sample std of C_T, and a bound. $L = 100$.



Errors and bounds for (Top) sparse sig., $N = 100, T = 5, K = 10$.
(Bottom): power-law decay signal, $N = 1000, T = 10$.

Near-sparse and noisy: simplified approach

Current solution $\hat{\mathbf{x}}$, true: $\mathbf{x}^*$. Take T new samples $y_i = \mathbf{a}_i'\mathbf{x}^*$, and compute $\hat{y}_i = \mathbf{a}_i'\hat{\mathbf{x}}$. Denote the error by $\delta = \hat{\mathbf{x}} - \mathbf{x}^*$, and let $z_i = \hat{y}_i - y_i$. Then

$$z_i = \mathbf{a}_i'\delta, \quad 1 \le i \le T$$

Now z_i's are i.i.d. from some a zero-mean distribution with variance $\|\delta\|_2^2 \, Var(a_{ij})$.

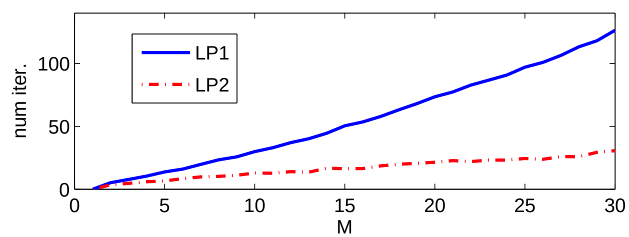
We can estimate $\|\delta\|_2^2$ by estimating the variance of the z_i. For example, for Gaussian $\mathbf{a}_i$, confidence bounds on $\|\delta\|_2^2$ can be obtained from the χ_T^2 distribution.

This is related to recent paper by Ward, "Compressed sensing with cross-validation" that uses the Johnson-Lindenstrauss lemma.

Solving sequential CS

Main goal of sequential CS – min number of samples. Yet, we also want efficient solution – not just resolving each time.

Warm-starting simplex: $\mathbf{x}^M$ is not feasible after $M+1$-st sample. Add a 'slack' variable: $\min \|\mathbf{x}\|_1 + Qz$, where $\mathbf{y}_{1:M} = A_M \mathbf{x}$, $y_{M+1} = \mathbf{a}'_{M+1}\mathbf{x} - z$, $z \geq 0$. For Q large enough, z is forced to 0.



Alternative approach: homotopy continuation between $\hat{\mathbf{x}}^M$ and $\hat{\mathbf{x}}^{M+1}$ – follow the piece-wise linear solution path. Garrigues and El Ghaoui, 2008, and indep. Asif and Romberg, 2008.

Summary and future work

Sequential processing can minimize the number of required measurements.

- Gaussian case: a simple rule requires the least possible number of samples.
- Bernoulli case: trade-off between probability of error and delay.
- Near-sparse and noisy case: Change in solutions gives us information about solution accuracy.

Interesting questions:

- Related results in graphical models structure recovery from samples, and low-rank matrix recovery.
- More efficient sequential solutions.
- Comparison with active learning approaches?